Stream Data Analisys and Data Mining With Twitter Storm

Gregor Majcen, Miha Zidar

***Abstract*—Twitter Storm is a powerfull distributed real time data processing solution, with a wide range of usage. In this paper, we are going to take a look at how we can utilize Twitter Storms power for data mining on streams of data. The main focus of this paper is real time data processing and online data mining.**

***Index Terms*—online learning, continuous data, data mining, distributed systems, horizontal scaling, batch processing**

I. Introduction

TODAY we're generating more information per second than ever before, and the amount of data produced is only increasing over time. Data on its own is not that useful for us, unless we can extract information from it and the speed of gathering that information is becomming more and more valuable. This is where the real time data processing comes in. Big companies like Twitter, Groupon, spider.io and others, are using Twitter Storm to provide a better user experience.

In the last few years data processing has come a long way with services like MapReduce, Amazon EMR, Hadoop, and related techologies. All of these were made to handle massive amounts of data, and they do that very affectively. But lately their weakness is showing in their lack of real time processing. Now speed is starting to be ever more important, to get ahead of competition. This is where Storm comes in. Unlike other solutions that work in ”batch” mode, Storm is made so it can handle data streams. With that it's possible to get calculations done on tupples and have the results back in matter of seconds, insted of hours or even days. Describe what the article contains.

We will begin by explaining how Twitter Storm works, and what it is used for. Then we shell slowly narow down all the Storm uses to the ones that become useful in datamining.

January 3, 2013

II. Twitter Storm

- spout
- bolt
- topology

A. Motivation

Before we can start talking about how Twitter Storm can be used, we need to first explain what it actually is and how it came to be.

B. Usage

- Stream processing -
- Distributed RPC
- Continous computation

C. Simple Example

```
N = [0] * n # initial reservoir
```

Explanation:

III. Online machine learning

A. Introduction

IV. Conclusion

Acknowledgement

The authors would like to thank Matjaž Kukar, PhD Assistant Professor.

References

[1] J. S. Vitter, *Random Sampling with a Reservoir*. Brown University, 1985.

[2] N. Littlestone, *Learning Quickly When Irrelevant Attributes Abound: A New Linear-threshold Algorithm*. University of California, 1988

[3] P. Zhao, R. Jing, *Online AUC Maximization* School of Computer Engineering, Nanyang Technological Unoversoty & Deparment of Computer Science and Engineering, Michigan State University

[4] R. Kessl, *Parallel algorithms for mining of frequent itemsets* The Faculty of Electrical Engineering, Czech Technical University in Prague

[5] Hanley, James A. and McNeil, Barbara J. *The meaning and use of the area under of receiver operating characteristic (roc) curve.* 1982.

Gregor Majcen 63070199

Miha Zidar 63060317